PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G10L 3/02 — A1

(11) International Publication Number: WO 98/19297

(43) International Publication Date: 7 May 1998 (07.05.98)

(21) International Application Number: PCT/US97/18761

(22) International Filing Date: 15 October 1997 (15.10.97)

(30) Priority Data:
08/739,975 30 October 1996 (30.10.96) US

(71) Applicant: MOTOROLA INC. [US/US]; 1303 East Algonquin Road, Schaumburg, IL 60196 (US).

(72) Inventors: CORRIGAN, Gerald; 1948 W. Farwell Avenue, Chicago, IL 60626 (US). KARAALI, Orhan; 5 Juniper, Rolling Meadows, IL 60008 (US). MASSEY, Noel; 4658 N. Sapphire Drive, Hoffman Estates, IL 60195 (US).

(74) Agents: STOCKLEY, Darleen, J. et al.; Motorola Inc., Intellectual Property Dept., 1303 East Algonquin Road, Schaumburg, IL 60196 (US).

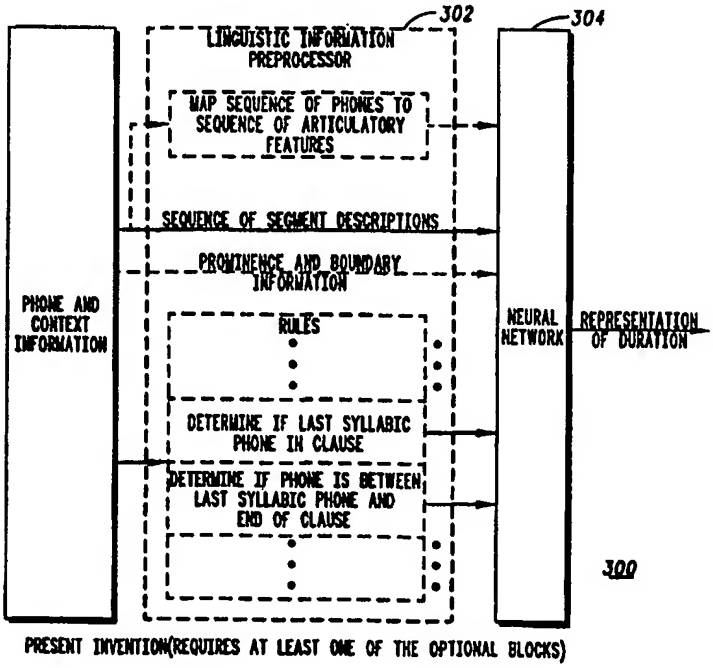
(81) Designated States: European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHOD, DEVICE AND SYSTEM FOR GENERATING SEGMENT DURATIONS IN A TEXT-TO-SPEECH SYSTEM

(57) Abstract

The present invention teaches a method (400), device and system (300) utilizing at least one of: mapping a sequence of phones to a sequence of articulatory features and utilizing prominence and boundary information, in addition to a predetermined set of rules for type, phonetic context, syntactic and prosodic context for phones to provide a system that generates segment durations efficiently with a small training set.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

METHOD, DEVICE AND SYSTEM FOR GENERATING SEGMENT DURATIONS IN A TEXT-TO-SPEECH SYSTEM

Field of the Invention

The present invention is related to text-to-speech synthesis, and more particularly, to segment duration generation in text-to-speech synthesis.

Background

To convert text to speech, a stream of text is typically converted into a speech wave form. This process generally includes determining the timing of speech events from a phonetic representation of the text. Typically, this involves the determination of the durations of speech segments that are associated with some speech elements, typically phones or phonemes. That is, for purposes of generating the speech, the speech is considered as a sequence of segments during each of which, some particular phoneme or phone is being uttered. (A phone is a particular manner in which a phoneme or part of a phoneme may be uttered. For example, the 't' sound in English, may be represented in the synthesized speech as a single phone, which could be a flap, a glottal stop, a 't' closure, or a

't' release. Alternatively, it could be represented by two phones, a 't' closure followed by a 't' release.) Speech timing is established by determining the durations of these segments.

In the prior art, rule-based systems generate segment durations using predetermined formulas with parameters that are adjusted by rules that act in a manner determined by the context in which the phonetic segment occurs, along with the identity of the phone to be generated during the phonetic segment. Present neural network-based systems provide full phonetic context information to the neural network, making it easy for the network to memorize, rather than generalize, which leads to poor performance on any phone sequence other than one of those on which the system has been trained.

Thus, there is a need for a neural network system that avoids the effects when the neural network depends only on chance correlations in training data and instead provides efficient segment durations.

Brief Description of the Drawings

FIG. 1 is a block diagram of a neural network that determines segment duration as is known in the art.

FIG. 2 is a block diagram of a rule-based system for determining segment duration as is known in the art.

FIG. 3 is a block diagram of a device/system in accordance with the present invention.

FIG. 4 is a flow chart of one embodiment of steps of a method in accordance with the present invention.

FIG. 5 illustrates a text-to-speech synthesizer incorporating the method of the present invention.

FIG. 6 illustrates the method of the present invention being applied to generate a duration for a single segment using a linguistic description.

Detailed Description of a Preferred Embodiment

The present invention teaches utilizing at least one of: mapping a sequence of phones to a sequence of articulatory features and utilizing prominence and boundary information, in addition to a predetermined set of rules for type, phonetic context, syntactic and prosodic context for segments to provide provide a system that generates segment durations efficiently with a small training set.

FIG. 1, numeral 100, is a block diagram of a neural network that determines segment duration as is known in the art. The input provided to the network is a sequence of representations of phonemes (102), one of which is the current phoneme, i.e., the phoneme for the current segment, or the segment for which the duration is being determined. The other phonemes are the phonemes associated with the adjacent segments, i.e., the segments that occur in sequence with the current segment. The output of the neural network (104) is the duration (106) of the current segment. The network is trained by obtaining a database of speech, and dividing it into a sequence of segments. These segments, their durations, and their contexts then provide a set of exemplars for training the neural network using some training algorithm such as back-propagation of errors.

FIG. 2, numeral 200, is a block diagram of a rule-based system for determining segment duration as is known in the art. In this example, phone and context data (202) is input into the rule-based system. Typically, the rule-based system utilizes certain preselected rules such as (1) determining if a segment is a last segment expressing a syllabic phone in a clause (204) and (2) determining if a segment is between a last segment expressing a syllabic phone and an end of a clause (206), multiplexes (208, 210) the outputs from the bipolar

question to weight the outputs in accordance with a predetermined scheme and send the weighted outputs to multipliers (212, 214) that are coupled serially to receive output information. The phone and context data then is sent as phone information (216) and a stress flag that shows whether the phone is stressed (218) to a look-up table (220). The output of the look-up table is sent to another multiplier (222) serially coupled to receive outputs and to a summer (224) that is coupled to the multiplier (222). The summer (224) outputs the duration of the segment.

FIG. 3, numeral 300, is a block diagram of a device/system in accordance with the present invention. The device generates segment durations for input text in a text-to-speech system that generates a linguistic description of speech to be uttered including at least one segment description. The device includes a linguistic information preprocessor (302) and a pretrained neural network (304). The linguistic information preprocessor (302) is operably coupled to receive the linguistic description of speech to be uttered and is used for generating an information vector for each segment description in the linguistic description, wherein the information vector includes a description of a sequence of segments surrounding the described segment and descriptive information for a context associated with the segment. The pretrained neural network (304) is operably coupled to the

linguistic information preprocessor (302) and is used for generating a representation of the duration associated with the segment by the neural network.

Typically, the linguistic description of speech includes a sequence of phone identifications, and each segment of speech is the portion of speech in which one of the identified phones is expressed. Each segment description in this case includes at least the phone identification for the phone being expressed.

Descriptive information typically includes at least one of: A) articulatory features associated with each phone in the sequence of phones; B) locations of syllable, word and other syntactic and intonational boundaries; C) syllable strength information; D) descriptive information of a word type; and E) rule firing information, i.e., information that causes a rule to operate.

The representation of the duration is generally a logarithm of the duration. Where desired, the representation of the duration may be adjusted to provide a duration that is greater than a duration that the pretrained neural network has been trained to provide. Typically, the pretrained neural network is a feedforward neural network that has been trained using back-propagation of errors.

Training data for the pretrained network is generated by recording natural speech, partitioning the speech data into identified phones, marking any other syntactical intonational and stress information used in the device and processing into informational vectors and target output for the neural network.

The device of the present invention may be implemented, for example, in a text-to-speech synthesizer or any text-to-speech system.

FIG. 4, numeral 400, is a flow chart of one embodiment of steps of a method in accordance with the present invention. The method provides for generating segment durations in a text-to-speech system, for input text that generates a linguistic description of speech to be uttered including at least one segment description. The method includes the steps of: A) generating (402) an information vector for each segment description in the linguistic description, wherein the information vector includes a description of a sequence of segments surrounding the described segment and descriptive information for a context associated with the segment; B) providing (404) the information vector as input to a pretrained neural network; and C) generating (406) a representation of the duration associated with the segment by the neural network.

As in the device, the linguistic description of speech includes a sequence of phone identifications and each segment of speech is the portion of speech in which one of the identified phones is expressed. Each segment description in this case includes at least the phone identification for the phone being expressed.

As in the device, descriptive information includes at least one of: A) articulatory features associated with each phone in the sequence of phones; B) locations of syllable, word and other syntactic and intonational boundaries; C) syllable strength information; D) descriptive information of a word type; and E) rule firing information.

Representation of the duration is generally a logarithm of the duration, and where selected, may be adjusted to provide a duration that is greater than a duration that the pretrained neural network has been trained to provide (408). The pretrained neural network is typically a feedforward neural network that has been trained using back-propagation of errors. Training data is typically generated as described above.

FIG. 5, numeral 500, illustrates a text-to-speech synthesizer incorporating the method of the present invention. Input text is analyzed (502) to produce a string of phones

(504), which are grouped into syllables (506). Syllables, in turn, are grouped into words and types (508), which are grouped into phrases (510), which are grouped into clauses (512), which are grouped into sentences (514). Syllables have an indication associated with them indicating whether they are unstressed, have secondary stress in a word, or have the primary stress in the word that contains them. Words include information indicating whether they are function words (prepositions, pronouns, conjunctions, or articles) or content words (all other words). The method is then used to generate (516) durations (518) for segments associated with each of the phones in the sequence of phones. These durations, along with the result of the text analysis, are provided to a linguistics-to-acoustics unit (520), which generates a sequence of acoustic descriptions (522) of short speech frames (10 ms. frames in the preferred embodiment). This sequence of acoustic descriptions is provided to a waveform generator (524), which produces the speech signal (526).

FIG. 6, numeral 600, illustrates the method of the present invention being applied to generate a duration for a single segment using a linguistic description (602). A sequence of phone identifications (604) including the identification of the phone associated with the segment for which a duration is being generated are provided as input to the neural network (610). In the preferred embodiment, this is a sequence of five

phone identifications, centered on the phone associated with the segment, and each phone identification is a vector of binary values, with one of the binary values in the vector set to one and the other binary values set to zero. A similar sequence of phones is input to a phone-to-feature conversion block (606), providing a sequence of feature vectors (608) as input to the neural network (610).

In the preferred embodiment, the sequence of phones provided to the phone-to-feature conversion block is identical to the sequence of phones provided to the neural network. The feature vectors are binary vectors, each determined by one of the input phone identifications, with each binary value in the binary vector representing some fact about the identified phone; for example, a binary value might be set to one if and only if the phone is a vowel. For one more similar sequence of phones, a vector of information (612) is provided describing boundaries which fall on each phone, and the characteristics of the syllables and words containing each phone. Finally, a rule firing extraction unit (614) processes the input to the method to produce a binary vector (616) describing the phone and the context for the segment for which duration is being generated. Each of the binary values in the binary vector is set to one if and only if some statement about the segment and its context is true; for example, "The segment is the last segment associated with a syllabic phone in the clause containing the

segment." This binary vector (616) is also provided to the neural network . From all of this input, the neural network generates a value which represents the duration. In the preferred embodiment, the output of the neural network (value representing duration, 618) is provided to an antilogarithm function unit (620), which computes the actual duration (622) of the segment.

The steps of the method may be stored in a memory unit of a computer or alternatively, embodied in a tangible medium of /for a Digital Signal Processor, DSP, an Application Specific Integrated Circuit, ASIC, or a gate array.

The present invention may be embodied in other specific forms without departing from its spirit or essential characteristics. The described embodiments are to be considered in all respects only as illustrative and not restrictive. The scope of the invention is, therefore, indicated by the appended claims rather than by the foregoing description. All changes which come within the meaning and range of equivalency of the claims are to be embraced within their scope.

We claim:

1. A method for generating segment durations in a text-to-speech system, wherein, for input text that generates a linguistic description of speech to be uttered including at least one segment description, comprising the steps of:

1A) generating an information vector for each segment description in the linguistic description, wherein the information vector includes a description of a sequence of segments surrounding a described segment and descriptive information for a context associated with the described segment;

1B) providing the information vector as input to a pretrained neural network; and

1C) generating a representation of a duration associated with the described segment by a neural network.

2. The method of claim 1 wherein at least one of 2A-2C:

2A) the speech is described as a sequence of phone identifications; the segments for which duration is being generated are segments of speech expressing predetermined phones in the sequence of phone identifications; and segment descriptions include the phone identifications; and where selected, wherein the descriptive information includes at least one of 2A1-2A5:

2A1) articulatory features associated with each phone in the sequence of phones;

2A2) locations of syllable, word and other syntactic and intonational boundaries;

2A3) syllable strength information;

2A4) descriptive information of a word type; and

2A5) rule firing information;

2B) the representation of the duration is a logarithm of the duration; and

2C) the representation of the duration is adjusted to provide a duration that is greater than a duration that the pretrained neural network has been trained to provide.

3. The method of claim 1 wherein the pretrained neural network is a feedforward neural network, and where selected, wherein the pretrained neural network has been trained using back-propagation of errors, and where further selected, wherein training data for the pretrained network has been generated by recording natural speech, partitioning the speech data into segments associated with identified phones, marking any other syntactical intonational and stress information used in the method and processing into informational vectors and target output for the neural network.

4. The method of claim 1 wherein at least one of 4A-4D:

4A) the steps of the method are stored in a memory unit of a computer;

4B) the steps of the method are embodied in a tangible medium of /for a Digital Signal Processor, DSP;

4C) the steps of the method are embodied in a tangible medium of/for an Application Specific Integrated Circuit, ASIC; and

4D) the steps of the method are embodied in a tangible medium of a gate array.

5. A device for generating segment durations in a text-to-speech system, for input text that generates a linguistic description of speech to be uttered including at least one segment description, comprising :

5A) a linguistic information preprocessor, operably coupled to receive the linguistic description of speech to be uttered, for generating an information vector for each segment description in the linguistic description, wherein the information vector includes a description of a sequence of segments surrounding a described segment and descriptive information for a context associated with a phoneme; and

5B) a pretrained neural network, operably coupled to the linguistic information preprocessor, for generating a representation of a duration associated with the described segment by the pretrained neural network.

6. The device of claim 5 wherein at least one of 6A-6D:

6A) the speech is described as a sequence of phone identifications; the segments for which the duration is being generated are segments of speech expressing predetermined phones in the sequence of phone identifications; and segment descriptions include the phone identifications, and where selected, wherein the descriptive information includes at least one of 6A1-6A5:

6A1) articulatory features associated with each phone in the sequence of phones;

6A2) locations of syllable, word and other syntactic and intonational boundaries;

6A3) syllable strength information;

6A4) descriptive information of a word type; and

6A5) rule firing information;

6B) the representation of the duration is a logarithm of the duration;

6C) the representation of the duration is adjusted to provide a duration that is greater than a duration that the pretrained neural network has been trained to provide; and

6D) the pretrained neural network is a feedforward neural network.

7. The device of claim 6 wherein, in 6D, the pretrained neural network has been trained using back-propagation of errors, and where selected, wherein training data for the pretrained network has been generated by recording natural

speech, partitioning speech data into segments associated with identified phones, marking any other syntactical intonational and stress information used in the device and processing into informational vectors and target output for the neural network.

8. A text-to-speech synthesizer having a device for generating segment durations in a text-to-speech system, for input text that generates a linguistic description of speech to be uttered including at least one segment description, the device comprising :

8A) a linguistic information preprocessor, operably coupled to receive the linguistic description of speech to be uttered, for generating an information vector for each segment description in the linguistic description, wherein the information vector includes a description of a sequence of segments surrounding a described segment and descriptive information for a context associated with a phoneme; and

8B) a pretrained neural network, operably coupled to the linguistic information preprocessor, for generating a representation of a duration associated with the described segment by the pretrained neural network.

9. The text-to-speech synthesizer of claim 8 wherein at least one of 9A-9D:

9A) the speech is described as a sequence of phone identifications; the segments for which duration is being generated are segments of speech expressing predetermined phones in the sequence of phone identifications; and segment descriptions include the phone identifications, and where selected, the information vector for each segment description includes at least one of 9A1-9A5:

9A1) articulatory features associated with each phone in the sequence of phones;

9A2) locations of syllable, word and other syntactic and intonational boundaries;

9A3) syllable strength information;

9A4) descriptive information of a word type; and

9A5) rule firing information;

9B) the representation of the duration is a logarithm of the duration;

9C) the representation of the duration is adjusted to provide a duration that is greater than a duration that the pretrained neural network has been trained to provide; and

9D) the pretrained neural network is a feedforward neural network.

10. The text-to-speech synthesizer of claim 9 wherein at least one of 10A-10B:

10A) the pretrained neural network has been trained using back-propagation of errors; and

10B) training data for the pretrained network has been generated by recording natural speech, partitioning the speech data into segments associated with identified phones, marking any other syntactical intonational and stress information used in the text-to-speech synthesizer and processing into informational vectors and target output for the neural network.

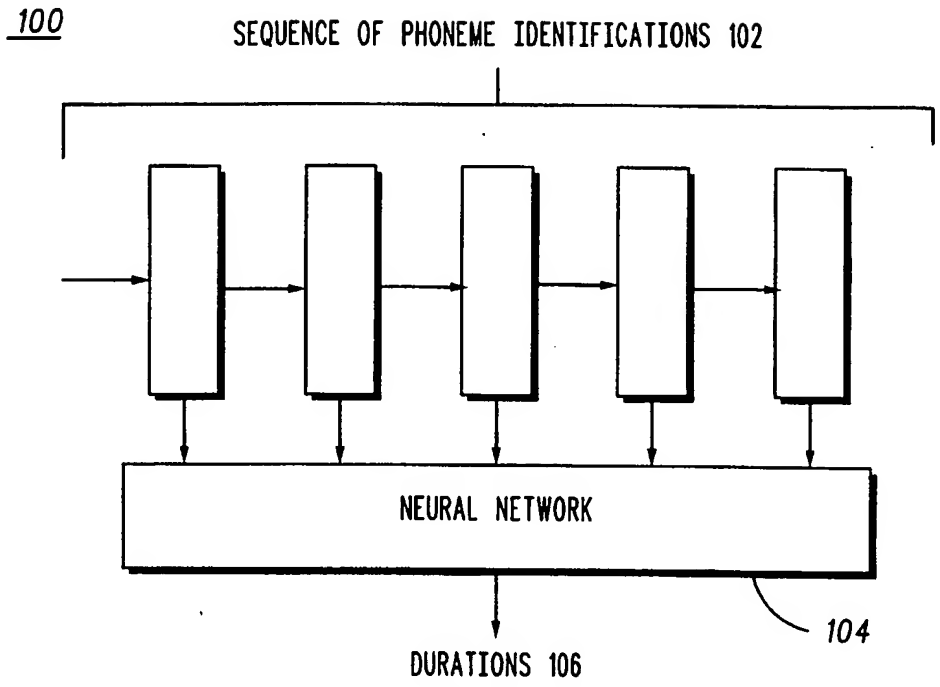

100
SEQUENCE OF PHONEME IDENTIFICATIONS 102
NEURAL NETWORK
104
DURATIONS 106

FIG.1

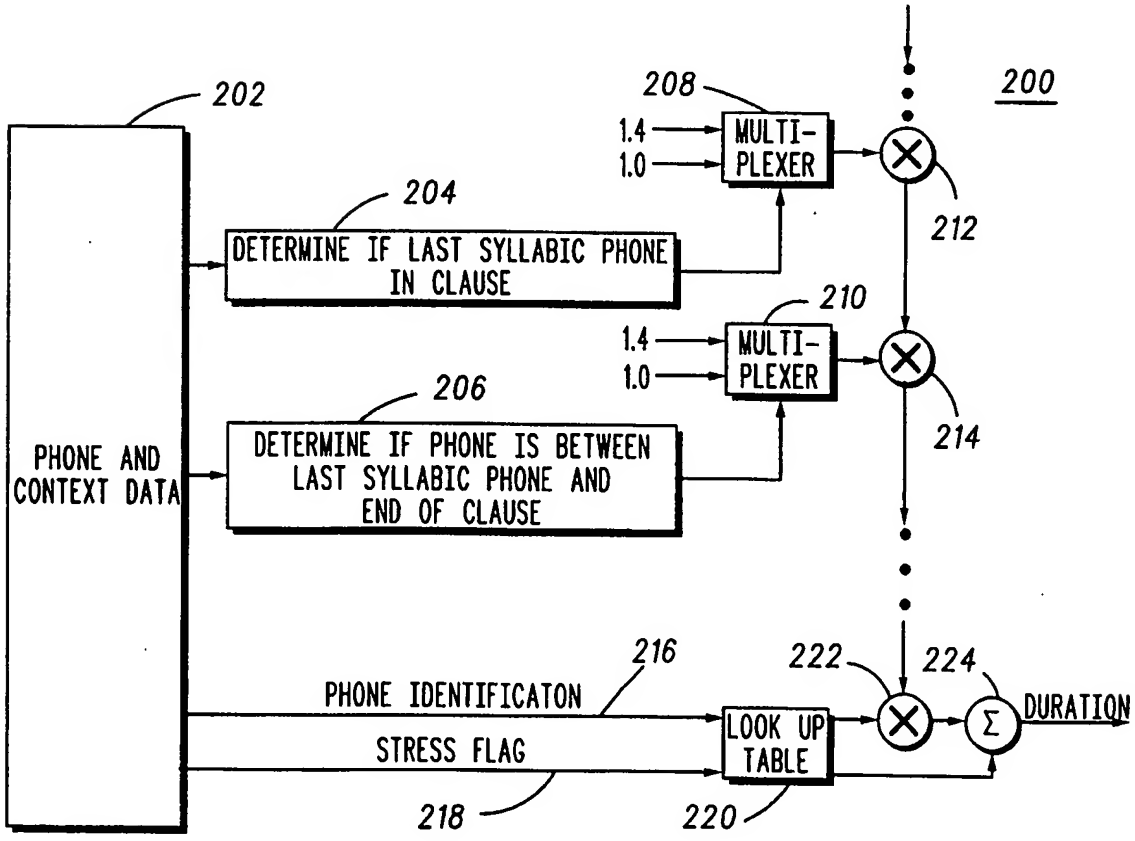
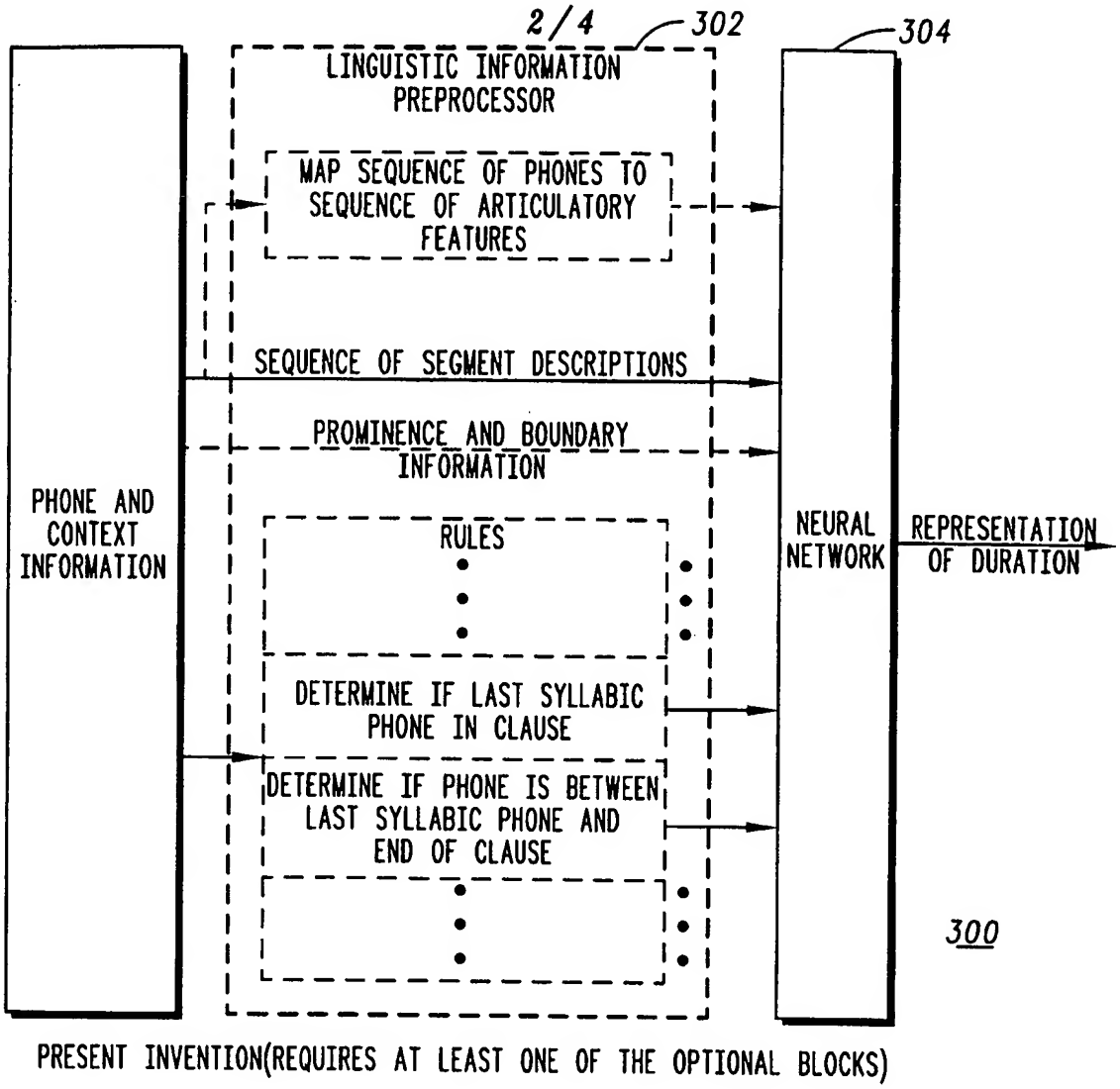

200
202
208
1.4
1.0
MULTI-PLEXER
212
204
DETERMINE IF LAST SYLLABIC PHONE IN CLAUSE
210
1.4
1.0
MULTI-PLEXER
214
206
DETERMINE IF PHONE IS BETWEEN LAST SYLLABIC PHONE AND END OF CLAUSE
PHONE AND CONTEXT DATA
216
222
224
PHONE IDENTIFICATON
STRESS FLAG
LOOK UP TABLE
Σ
DURATION
218
220

FIG.2



PRESENT INVENTION(REQUIRES AT LEAST ONE OF THE OPTIONAL BLOCKS)

FIG.3

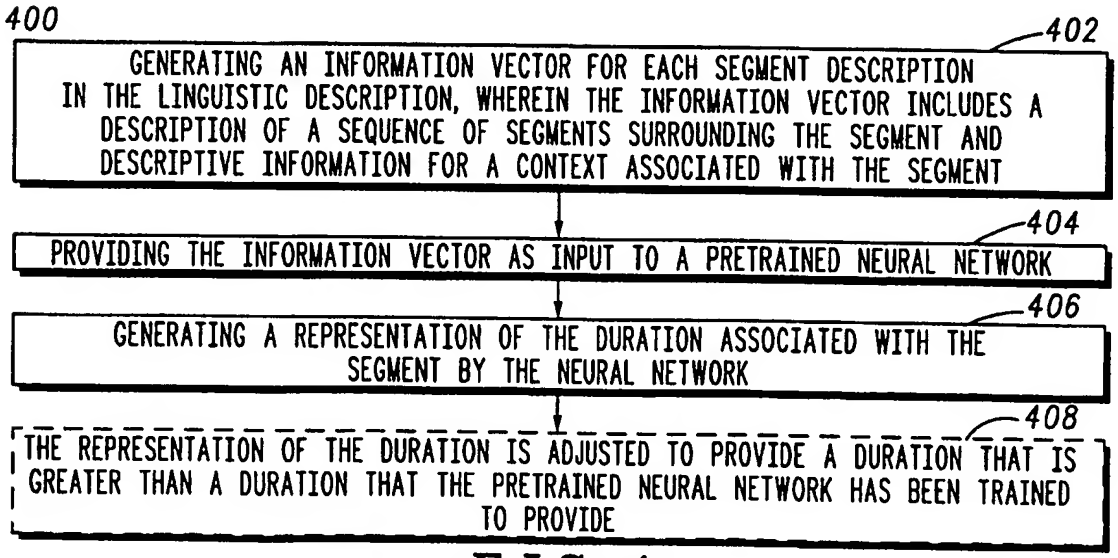


FIG.4

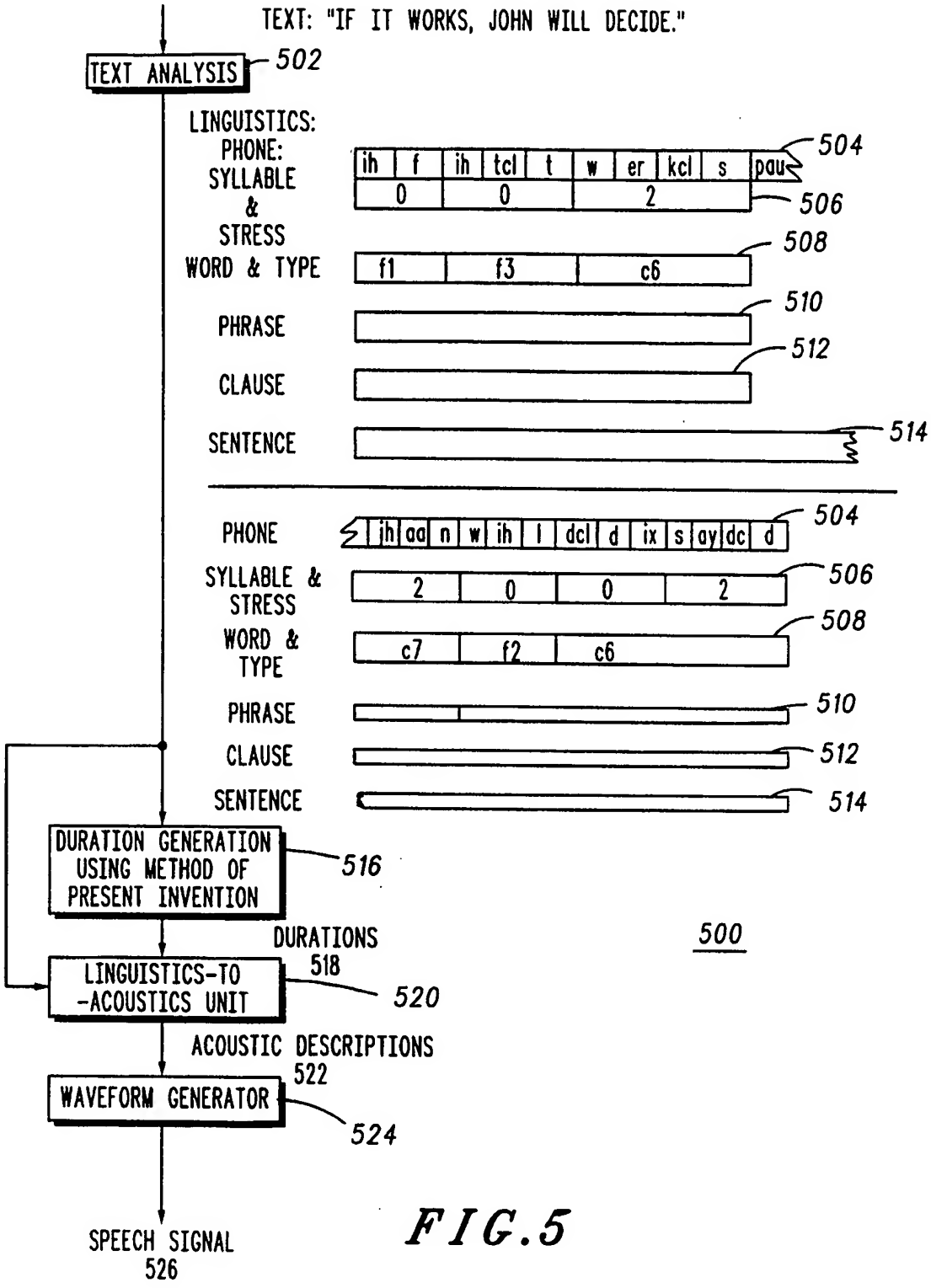

TEXT: "IF IT WORKS, JOHN WILL DECIDE."
TEXT ANALYSIS
502
LINGUISTICS:
PHONE:
SYLLABLE & STRESS
WORD & TYPE
PHRASE
CLAUSE
SENTENCE
ih f ih tcl t w er kcl s pau
504
0 0 2
506
f1 f3 c6
508
510
512
514
PHONE
SYLLABLE & STRESS
WORD & TYPE
PHRASE
CLAUSE
SENTENCE
ih aa n w ih l dcl d ix s ay dc d
504
2 0 0 2
506
c7 f2 c6
508
510
512
514
DURATION GENERATION USING METHOD OF PRESENT INVENTION
516
DURATIONS
518
LINGUISTICS-TO-ACOUSTICS UNIT
520
ACOUSTIC DESCRIPTIONS
522
WAVEFORM GENERATOR
524
SPEECH SIGNAL
526
500

FIG.5

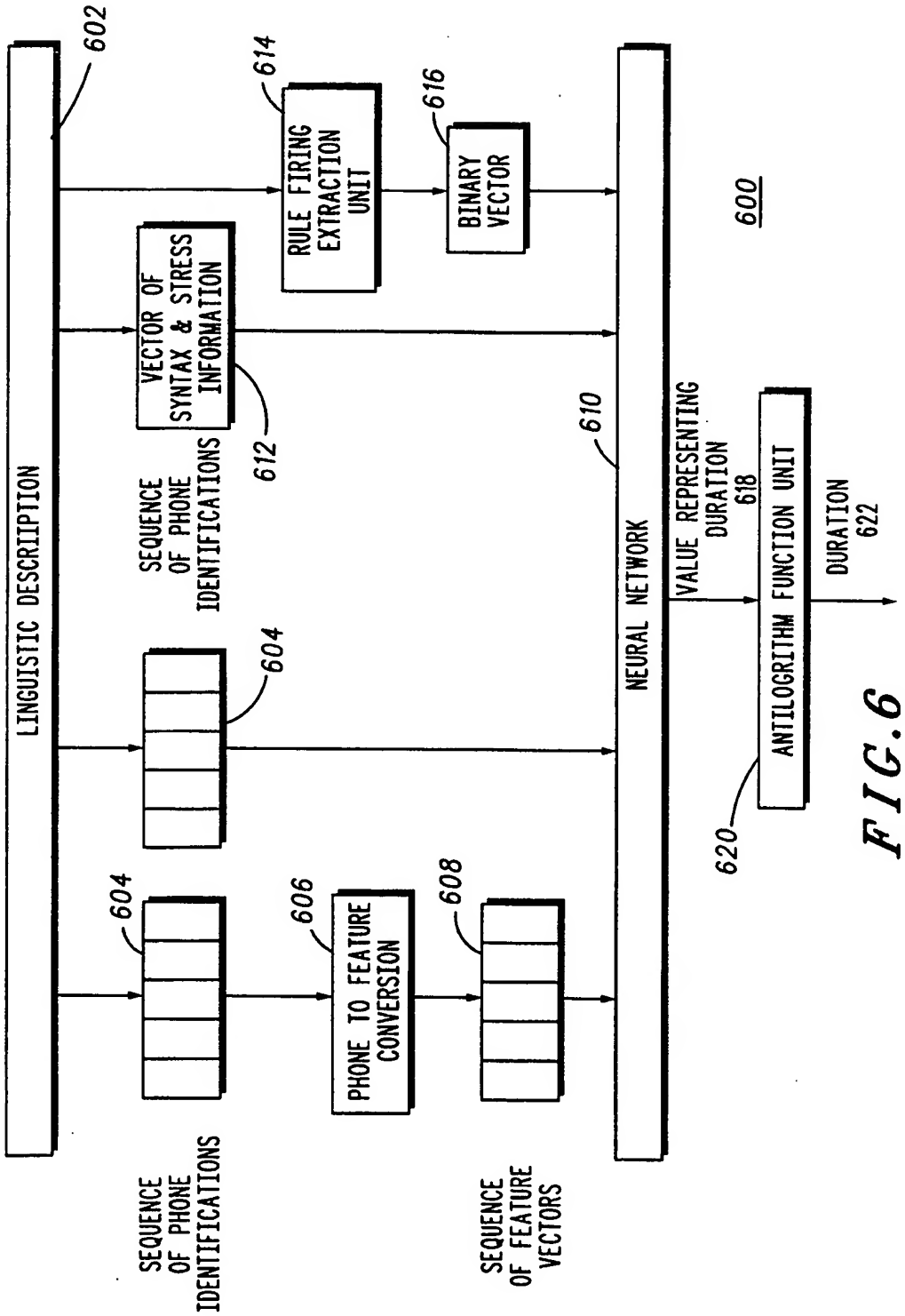

LINGUISTIC DESCRIPTION
602
VECTOR OF SYNTAX & STRESS INFORMATION
612
RULE FIRING EXTRACTION UNIT
614
BINARY VECTOR
616
600
SEQUENCE OF PHONE IDENTIFICATIONS
604
NEURAL NETWORK
610
VALUE REPRESENTING DURATION
618
ANTILOGRITHM FUNCTION UNIT
620
DURATION
622
SEQUENCE OF PHONE IDENTIFICATIONS
604
PHONE TO FEATURE CONVERSION
606
SEQUENCE OF FEATURE VECTORS
608

FIG. 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US97/18761

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :G10L 3/02
US CL : 704/260, 259
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 704/260, 259, 232, 200, 268

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5,463,713 A (HASEGAWA) 31 October 1995; Fig. 2, Fig. 3; Fig. 4; col. 2, lines 10-30 and col. 4, line 56 to col. 5, line 21.	1-10
Y	US 5,327,498 A (HAMON) 05 July 1994; Fig. 1; Fig. 3; col. 2; col. 3; and col. 7	1-10
Y	US 5,230,037 A (GIUSTINIANI ET AL.) 20 July 1993; col. 9, lines 23-56.	2 and 6
A, P	US 5,610,812 A (SCHABES ET AL.) 11 March 1997; col. 4; col. 5; col. 6.	1-10

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search	Date of mailing of the international search report
09 JANUARY 1998	25/02/1998
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer SCOTT RICHARDSON Telephone No. (703) 305-3900

Form PCT/ISA/210 (second sheet)(July 1992)*